Impact of choice of volumetry software and nodule management guidelines on recall rates in lung cancer screening

Abstract

Purpose

Appropriate lung nodule management is essential to minimizing unnecessary patient recall in lung cancer screening. Two European guidelines provide differing recommendations in that participants with nodules ≥100mm^3 or ≥80mm^3 respectively should be recalled, at baseline. Nodule size estimation is known to vary between volumetry software packages (VSPs). The aim of this study was to examine the impact of choice of VSP on participant recall rates, when applying different European nodule management guidelines. An additional aim was to compare recall rates between 7 VSPs and manual diameter measurements.

Methods

156 small-sized lung nodules (50-150mm^3) from the UK Lung Screening trial were measured using 7 different VSPs (VSP1-7) and also using manual diameter. The type of VSP used in the NELSON study (VSP1), on which European nodule management guidelines are based, provided the reference standard. Nodule size was compared using Bland Altman, and recall rates by Mcnemar's test.

Results

Compared to the reference standard, a 100mm^3 threshold for recall, resulted in no difference in recall rates only for VSP 5 & 7. Using an 80mm^3 threshold resulted in no difference in recall rates for VSP2 & 6. Recall rates were significantly higher for VSP 4 regardless of threshold and when using manual diameter measurements.

Conclusions

Appropriate nodule size thresholds for recall in screening depend on the type of volumetry software used. The results highlight the importance of benchmarking of volumetry packages

Key words

Lung cancer screening, pulmonary nodule, volumetry; computed tomography

Abbreviations

CT – Computed tomography

VSP – Volume software package

BTS – British Thoracic Society

UKLS – United Kingdom Lung Cancer screening

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

1. Introduction

Following the National Lung Screening Trial (NLST) [1] and more recently presented results from the NELSON trial[2] showing a mortality reduction from lung cancer screening, many European countries are deciding whether to implement lung cancer screening with CT. A critical factor in this decision will be cost effectiveness, and central to this is the appropriate minimization of false positives. Although various definitions of "false positive" have been used in the CT screening literature, a practical definition is any screening result that leads to a participant being recalled for further tests, including repeat CT, prior to the next scheduled screening round. This has also been referred to as the "recall rate" [3, 4].

Evidence from a number of studies has shown that the use of nodule volumetry software and the use of higher nodule size thresholds for repeat CT compared to that used in early lung cancer screening trials can reduce recall rates [5]. Based on this evidence, a recent position paper by European lung cancer screening investigators recommended that volumetry should be the preferred method of nodule size measurement [6]. It also recommended that only participants with solid nodules ≥ 100mm^3 in volume at baseline should be recalled for repeat CT, while participants with solid nodules < 100 mm^3 did not require CT before their next scheduled screening round. This was based on data from the NELSON lung cancer screening trial which demonstrated that participants with nodules ≥100mm^3 had a significantly increased risk of developing lung cancer [7]. That evidence was generated using one particular nodule volumetry software package.

It is known from a small number of studies that there is a variation in nodule size estimation between different volumetry software packages (VSPs)[8-10]. Therefore, to account for the fact that screening programmes would not necessarily have access to the same volumetry sofware used in the NELSON study, the British Thoracic Society (BTS) guidelines chose a more cautious cut-off and recommended that patients with nodules on CT <80mm^3 (instead of 100mm^3) did not require early recall [11].

Previous studies have not examined the degree to which the choice of nodule VSP and choice of nodule volumetry management guideline in combination, influences recall rates in lung cancer screening. Therefore, in this study, we aimed to evaluate the impact on recall rates in a CT screening population enriched for small sized nodules, using 7 different commercially available VSPs currently in use in radiology departments, (including the type of VSP used in the NELSON lung cancer screening trial), applying nodule size thresholds for recall of 80mm^3 and 100mm^3 respectively.

Since volumetry is not universally available and because volumetry is not usable on all nodules, nodule size may sometimes need to be measured using electronic callipers. In this scenario, the European position statement recommends that a threshold of ≥5mm nodule diameter should be employed before participants are recalled for repeat CT [6]. Therefore, an additional aim of this study was to compare recall rates between 7 VSPs and manual diameter measurements.

2. Methods and materials

The study had ethics approval from the National Research Ethics Service and patient consent was obtained.

2.1 Patient Cohort

The study population was chosen from the United Kingdom Lung Cancer Screening (UKLS) study baseline nodule database, derived from the 1,994 participants randomised to the CT intervention arm of the study [4]. Since the purpose of the current study was to examine the impact on recall rates when applying either a 80mm^3 or 100mm^3 cut-off, a cohort of small sized nodules centred around this threshold was deliberately chosen for the study. All solid nodules between 50mm^3 and 150mm^3 classified as requiring recall at the time of the UKLS study, by UKLS thoracic radiologists, at consecutive baseline CTs from the UKLS study were identified from the UKLS database (n=173 nodules in 131 patients). (Benign nodules including those regarded as calcified or intrapulmonary

lymph nodes by UKLS radiologists did not require recall in the UKLS study and therefore are not included in the current study). 17 nodules were excluded from the analysis because corruption of CT data that occurred during image transfer did not allow for analysis using all volumetry packages, leaving 156 nodules in 119 subjects in the final dataset.

2.2 Image Acquisition

Non-contrast thoracic CT images were obtained craniocaudally in suspended maximal inspiration. Axial images were reconstructed at 1-mm thickness with 0.7 increments using a moderate spatial frequency kernel. Further details on the UKLS screening trial CT and reading protocol is provided elsewhere [4].

2.3 Image Analysis

All nodules were independently evaluated by one of six radiologists, (ranging between five and fourteen years' experience), on dedicated workstations containing commercially available volumetry applications currently in use in radiology departments. Because the thresholds for nodule management in the European position statement [6] and British Thoracic Society Guidelines [11] were derived from data from the NELSON trial [12], the type of VSP used in the NELSON trial [Syngo MMWP LungCare VB10A] is regarded as the reference standard for this study, and referred to as VSP 1.

The other VSPs tested were Siemens SyngoVia MM Oncology version VB10B; FUJI Synapse 3d version 4.4EU; GE AW 3.2 LungVCAR; Philips Lung Nodule Application (LNA) Extended Brilliance Workspace version 4.5.6.52040; Terarecon Aquarious iNtuitiion 4.4.12; and Vitrea Vital images v.6.2. As per precedence from previous studies examining nodule software reliability [8, 10] the results for these VSPs are blinded and referred to as VSP 2-7 in the results below.

To ensure consistency, each radiologist was provided with the axial slice position and segmental anatomical location reference for each nodule, as documented in the UKLS database. The volume of each nodule was measured according to the specific VSP's manufacturer's instructions. This was either initiated by the reader clicking the nodule centre, or drawing a line across the nodule. Manual alteration of the segmented region was not permitted. The reliability of automated segmentation was subjectively assessed by each reader based on the criteria described by de Hoop et al. [8]. That study considered a nodule to be unreliably segmented when the segmented nodule volume boundaries were estimated to exceed approximately 30% of the nodule, judged visually. Typically, this would occur when the nodule segmentation included adjacent structures such as a pulmonary vessel. Nodules that were not reliably segmented were recorded as such.

A separate analysis was performed by two further thoracic radiologists (with five and seven years' experience) to manually measure the long axis diameter of the same nodule dataset on axial CT. Diameters were measured using electronic callipers to one decimal point on a lung window setting (W1500, L -500).

2.4 Statistical Methods

Nodule volume and diameter were described by median and range. Nodule size using six semi-automated volumetry software packages (VSP2-7) was compared to nodule size obtained from the reference standard, VSP 1, using Bland Altman statistics to illustrate bias and 95% limits of agreement [13]. For the purposes of the study, recall rate was determined on a per nodule basis, i.e. each nodule was regarded as a separate case. The proportion of nodules requiring recall using VSP2-7 or manual diameter compared to VSP1 was evaluated using Mcnemar's test. P values < 0.05 were regarded as significant. Only nodules that were judged to be adequately segmented by VSP 1 and the comparative measurement method were evaluated for each analysis. All statistical analyses were performed using Medcalc version 7.4 software.

3. Results

3.1 Nodule data

The median and range nodule size for the 156 nodules acquired from the seven VSPs is provided in Table 1. The number and proportion of nodules that could not be segmented reliably per VSP is also provided in Table 1. Table 1 also shows median and range nodule diameter measured by the two readers. The distribution of the nodules was as follows: right upper lobe 40, right middle lobe 16, right lower lobe 30, left upper lobe 17, lingula 7 and left lower lobe 46.

Compared to VSP 1; VSP 4 and VSP 5 overestimated nodule volume, whereas VSP 2, 3, 6 and 7 underestimated nodule volume. Nodule volume comparison between VSP 2-7 with their respective 95% limits of agreement and VSP 1 are provided in Table 2 and visually represented in Bland Altman plots, Figures 1.1 to 1.6.

3.2 Influence of nodule volumetry software packages and manual diameter measurement on recall rates, when applying European Nodule Management Recommendations

When using a 100mm^3 threshold, the recall rate in this cohort enriched for small nodules ranged from 10.0% for VSP 6 to 94.3% for VSP 4 (Table 3). Three VSPs (VSP 2,3 and 6) had significantly lower recall rates compared to VSP 1. VSP 5 and VSP 7 showed no difference in recall rates, whereas VSP 4 was associated with significantly higher recall rates. The use of manual diameter measurements and a 5mm threshold also led to significantly greater recalls for both Reader 1 and Reader 2 compared to VSP 1.

Using an 80mm^3 threshold for VSPs 2-7 resulted in significantly higher recall rates for VSP 3, 4, 5 and 7 compared to VSP 1 (Table 4). VSP 2 and VSP 6 no longer demonstrated significantly different recall rates compared to VSP 1.

4. Discussion

The results from our study suggest that the most appropriate lung nodule size threshold for participant recall at baseline lung cancer screening CT should depend on the type of nodule volumetry software used. Our study shows that for two of the tested VSPs (VSP 5 and 7), recall rates matched the performance of the type of VSP used in the NELSON trial (VSP 1), indicating that the 100mm^3 threshold for recall recommended in the European Lung Cancer Screening Position Statement is appropriate.

The BTS nodule management guidelines deliberately recommended a lower, more cautious threshold of 80mm^3 for repeat CT, to take into account possible variability in nodule size estimation between VSPs. Our results show that for two other VSPs (VSP 2 and VSP 6), that on average underestimated nodules size compared to VSP 1, an 80mm^3 cut-off for repeat CT was more appropriate than 100mm^3.

We also showed that for one VSP (VSP 4), the degree of size overestimation was substantial, such that recall rates exceeded 90% regardless of nodule size threshold. However, this particular VSP was an outlier. All of the five remaining VSPs had ≤7% difference in recall rates compared to VSP 1, when using either an 80mm^3 or 100mm^3 threshold. To put this into context, recall rates were 50.7% - 59.4% when using manual diameter measurements, highlighting the limitations of electronic calliper measurements to accurately measure nodule size. The greater tendency towards unnecessary follow-up when using diameter compared to the majority of VSPs is important in the context of lung cancer screening implementation methodology, where the burden of false positives is cited as a barrier to the policy [14].

The degree of variation in nodule volume measurements between VSPs was larger in our study than that previously reported by Zhao et al. [10] and de Hoop et al. [8]. This could potentially be explained by the fact that we tested a greater number of VSPs than previous studies. Indeed, a

strength of our study is that we examined a wide range of VSPs currently in use in radiology departments, and which might theoretically be used in lung screening programmes today. Results from the study by de Hoop et al [8] for example, (from 2009), evaluated volumetry software versions no longer typically in use. However, it is acknowledged that there are other VSPs currently in use in European radiology departments, outside of those tested in our study. Until formal performance benchmarking across all vendors is performed, it could be suggested that future screening programmes may wish to choose the more cautious threshold of 80mm^3 for recall. This may be particularly appropriate if, for cost-effectiveness reasons, countries decide to implement biennial screening [6] as participants would only undergo a subsequent CT after two years.

The main limitation of this study is that our results were benchmarked against the type of VSP used in the NELSON study and not based upon a true histopathological proven outcomes; therefore it is possible that some nodules labelled as "unnecessary recall" may have turned out to be malignant in time. However, we believe that it is appropriate to use VSP 1 as the reference standard, as it has previously been demonstrated to be a reliable surrogate of outcome in a study of >9000 nodules from the NELSON trial [7], and is the basis upon which European nodule management guidelines have been based. Furthermore, our results capture what is important in clinical practice, which is whether a nodule is actionable based on current guidelines. Second, we only evaluated one aspect of volumetry, and did not examine the reliability of volume doubling time calculation which is another important feature when using volumetry. We also did not compare intra-package variability, although this is more relevant for the reliability of evaluation of nodule size over time and has also previously been shown to be minimal [15]. We also focused on solid nodules and did not evaluate subsolid nodules or potentially sinister morphological characteristics that would usually be assessed in a clinical context. That said for the smaller sized nodules examined in this study it is likely that such differentiating characteristics are less meaningful.

In conclusion our study is, to the best of our knowledge, the first to examine the reliability of a wide range of commercially available nodule volumetry packages currently in use in radiology departments, and the corresponding impact on screening recall rates when applying different European lung nodule management recommendations. For screening to be successfully implemented, clinicians need to know that the VSP used in their screening programme is comparable to those used to set guidelines. Further work is need to collate standardised nodule datasets, such that all VSPs can be tested for performance and consistency via formal benchmarking exercises.

References

[1] D.R. Aberle, A.M. Adams, C.D. Berg, W.C. Black, J.D. Clapp, R.M. Fagerstrom, I.F. Gareen, C. Gatsonis, P.M. Marcus, J.D. Sicks, Reduced lung-cancer mortality with low-dose computed tomographic screening, N.Engl.J.Med. 365(5) (2011) 395-409.
[2] H. De Koning, Van Der Aalst C., Ten Haaf K., Oudkerk M., Effects of Volume CT Lung Cancer Screening: Mortality Results of the NELSON Randomised-Controlled Popluation Based Trial, Journal of Thoracic Oncology 13(10) (2018) S185.
[3] N. Becker, E. Motsch, M.L. Gross, A. Eigentopf, C.P. Heussel, H. Dienemann, P.A. Schnabel, M. Eichinger, D.E. Optazaite, M. Puderbach, M. Wielputz, H.U. Kauczor, J. Tremper, S. Delorme, Randomized Study on Early Detection of Lung Cancer with MSCT in Germany: Results of the First 3 Years of Follow-up After Randomization, J Thorac Oncol 10(6) (2015) 890-6.
[4] J.K. Field, S.W. Duffy, D.R. Baldwin, D.K. Whynes, A. Devaraj, K.E. Brain, T. Eisen, J. Gosney, B.A. Green, J.A. Holemans, T. Kavanagh, K.M. Kerr, M. Ledson, K.J. Lifford, F.E. McRonald, A. Nair, R.D. Page, M.K. Parmar, D.M. Rassl, R.C. Rintoul, N.J. Screaton, N.J. Wald, D. Weller, P.R. Williamson, G. Yadegarfar, D.M. Hansell, UK Lung Cancer RCT Pilot Screening Trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening, Thorax 71(2) (2016) 161-70.
[5] R. Yip, C.I. Henschke, D.F. Yankelevitz, J.P. Smith, CT screening for lung cancer: alternative definitions of positive test result based on the national lung screening trial and international early lung cancer action program databases, Radiology 273(2) (2014) 591-6.
[6] M. Oudkerk, A. Devaraj, R. Vliegenthart, T. Henzler, H. Prosch, C.P. Heussel, G. Bastarrika, N. Sverzellati, M. Mascalchi, S. Delorme, D.R. Baldwin, M.E. Callister, N. Becker, M.A. Heuvelmans, W. Rzyman, M.V. Infante, U. Pastorino, J.H. Pedersen, E. Paci, S.W. Duffy, H. de Koning, J.K. Field, European position statement on lung cancer screening, Lancet Oncol 18(12) (2017) e754-e766.
[7] N. Horeweg, R.J. van, M.A. Heuvelmans, C.M. van der Aalst, R. Vliegenthart, E.T. Scholten, H.K. Ten, K. Nackaerts, J.W. Lammers, C. Weenink, H.J. Groen, O.P. van, P.A. de Jong, G.H. de Bock, W. Mali, H.J. de Koning, M. Oudkerk, Lung cancer probability in patients with CT-detected pulmonary nodules: a prespecified analysis of data from the NELSON trial of low-dose CT screening, Lancet Oncol. 15(12) (2014) 1332-1341.
[8] H.B. de, H. Gietema, G.B. van, P. Zanen, G. Groenewegen, M. Prokop, A comparison of six software packages for evaluation of solid lung nodules using semi-automated volumetry: what is the minimum increase in size to detect growth in repeated CT examinations, Eur.Radiol. 19(4) (2009) 800-808.
[9] M. Liang, R. Yip, W. Tang, D. Xu, A. Reeves, C.I. Henschke, D.F. Yankelevitz, Variation in Screening CT-Detected Nodule Volumetry as a Function of Size, AJR Am J Roentgenol 209(2) (2017) 304-308.
[10] Y.R. Zhao, P.M. van Ooijen, M.D. Dorrius, M. Heuvelmans, G.H. de Bock, R. Vliegenthart, M. Oudkerk, Comparison of three software systems for semi-automatic volumetry of pulmonary nodules on baseline and follow-up CT examinations, Acta Radiol. 55(6) (2014) 691-698.
[11] M.E. Callister, D.R. Baldwin, A.R. Akram, S. Barnard, P. Cane, J. Draffan, K. Franks, F. Gleeson, R. Graham, P. Malhotra, M. Prokop, K. Rodger, M. Subesinghe, D. Waller, I. Woolhouse, G. British Thoracic Society Pulmonary Nodule Guideline Development, C. British Thoracic Society Standards of Care, British Thoracic Society guidelines for the investigation and management of pulmonary nodules, Thorax 70 Suppl 2 (2015) ii1-ii54.
[12] R.J. van Klaveren, M. Oudkerk, M. Prokop, E.T. Scholten, K. Nackaerts, R. Vernhout, C.A. van Iersel, K.A. van den Bergh, W.S. van 't, A.C. van der, E. Thunnissen, D.M. Xu, Y. Wang, Y. Zhao, H.A. Gietema, B.J. de Hoop, H.J. Groen, G.H. de Bock, O.P. van, C. Weenink, J. Verschakelen, J.W. Lammers, W. Timens, D. Willebrand, A. Vink, W. Mali, H.J. de Koning, Management of lung nodules detected by volume CT scanning, N.Engl.J.Med. 361(23) (2009) 2221-2229.
[13] J.M. Bland, D.G. Altman, Statistical methods for assessing agreement between two methods of clinical measurement, Lancet 1(8476) (1986) 307-10.

[14] M.S. Fabrikant, J.P. Wisnivesky, T. Marron, E. Taioli, R.R. Veluswamy, Benefits and Challenges of Lung Cancer Screening in Older Adults, Clin Ther 40(4) (2018) 526-534.
[15] Y. Wang, R.J. van Klaveren, H.J. van der Zaag-Loonen, G.H. de Bock, H.A. Gietema, D.M. Xu, A.L. Leusveld, H.J. de Koning, E.T. Scholten, J. Verschakelen, M. Prokop, M. Oudkerk, Effect of nodule characteristics on variability of semiautomated volume measurements in pulmonary nodules detected in a lung cancer screening program, Radiology 248(2) (2008) 625-631.

Tables

Table 1: Nodule size and segmentation reliability

	VSP 1	VSP 2	VSP 3	VSP 4	VSP 5	VSP 6	VSP7	Manual Diameter Reader 1	Manual Diameter Reader 2
Number (%) of nodules not reliably segmented	18 (12)	27 (17)	11 (7)	41(26)	22(14)	33 (21)	37 (24)	N/A	N/A
Median (Range) nodule size [mm^3 for VSP1-7, mm for Readers 1 & 2]	70.0 (50.1-147.8)	64.9 (40.2-157.0)	60.0 (18.0-163.0)	178.0 (68.0-596.0)	78.6 (22.3-385.9)	52.3 (21.2-334.7)	68.0 (26.0-149.0	4.9 (2.5 - 8.3)	5.1 (2.9 - 8.6)

Nodule size and segmentation reliability of the VSPs 1-7, and for two readers using manual diameter. N/A= not applicable.

Table 2: Comparison of nodule volumes

	VSP 2	VSP 3	VSP 4	VSP 5	VSP 6	VSP 7
Mean % volume measurement variability	-12.3 (-57.7, 33.0)	-17.0 (-64.3, 30.3)	+77.0 (22.4, 131.6)	+10.5 (-37.7, 58.7)	-27.0 (-94.7, 40.6)	-5.1 (-47.7, 37.2)

Comparison of nodule volumes of VSP 2-7 compared to VSP 1, expressed as mean % volume measurement variability (95% limits of agreement). A negative volume measurement difference denotes an underestimation of nodule volume and a positive difference reflects an overestimation of nodule volume compared to the reference reading method VSP 1.

Table 3. Comparison of recall rates using a 100mm^3 or 5mm diameter threshold, based on European Investigators Position Statement Recommendations

Comparison reading method	No. of nodules	VSP 1 Recall Rate (%)	Comparison Recall rate (%)	Difference (%)	p-value
VSP 2	116	24.1	12.1	-12.0	**0.0013**
VSP 3	130	24.6	17.7	-6.9	**0.0225**
VSP 4	106	27.4	94.3	66.9	**<0.0001**
VSP 5	123	25.2	30.1	4.9	0.1460
VSP 6	110	20.0	10.0	-10.0	**0.0192**
VSP 7	111	26.1	23.4	-2.7	0.4531
Manual Diameter Reader 1	138	26.1	50.7	24.6	**<0.0001**
Manual Diameter Reader 2	138	26.1	59.4	33.3	**<0.0001**

Recall rates comparing 100mm^3 threshold for VSPs 2-7 and 5mm threshold for Readers 1 & 2 versus 100mm^3 threshold for VSP 1. A negative difference denotes a lower recall rate using the reading method as compared to the reference reading method VSP 1.

Table 4. Comparison of recall rates using a 80mm^3 threshold, based on British Thoracic Society Recommendations

Comparison Reading method	No. of nodules	VSP 1 Recall Rate (%)	Comparison Recall rate (%)	Difference (%)	p-value
VSP 2	116	24.1	26.7	2.6	0.6291
VSP 3	130	24.6	31.5	6.9	**0.0117**
VSP 4	106	27.4	99.1	71.7	**<0.0001**
VSP 5	123	25.2	50.4	25.2	**<0.0001**
VSP 6	110	20.0	19.1	-0.9	1.0
VSP 7	111	26.1	36.0	9.9	**0.0074**

Recall rates comparing 80mm^3 threshold for VSPs 2-7 versus 100mm3 threshold for VSP 1. A negative difference denotes a lower recall date using the reading method as compared to the reference reading method VSP 1.